ABSTRACT

The present invention minimizes the amount of traffic that traverses the fabric in support of the cache coherency protocol. It also allows rapid transmission of all traffic associated with the cache coherency protocol, so as to minimize latency and maximize performance. A fabric is used to interconnect a number of processing units together. The switches are able to recognize incoming traffic related to the cache coherency protocol and then move these messages to the head of that switch's output queue to insure fast transmission. Also, the traffic related to the cache coherency protocol can interrupt an outgoing message, further reducing latency. The switch incorporates a memory element, dedicated to the cache coherency protocol, which tracks the contents of all of the caches of all of the processors connected to the fabric. In this way, the fabric can selectively transmit traffic only to the processors where it is relevant.